INM431 Machine Learning – Comparative Study of two different Machine Learning Algorithms, Random Forest and K-Nearest Neighbours, for the purpose of forecasting Sales

CITY UNIVERSITY LONDON

Daniel Dixey and Enrico Lopedoto

MSc Data Science, School of Mathematics, Computer Science & Engineering, Department of Computer Science

Motivation and Description

Predictive analytics can be argued as of the most alluring and effective methods of forecasting the future given past observations from all businesses and organizations. To quote Barton and Court (2012) [**1**], “Advanced analytics is likely to become a decisive competitive asset in many industries and a core element in companies' efforts to improve performance.” Machine style solutions offer far more robust methods for this type of analysis, as they are able to take into account many variables (*high dimensionality*) with an accuracy that can far exceed the capability of any human, even one with specialist domain knowledge of the particular area.

The aim of this analysis is to explore, compare and contrast two Machine Learning algorithms for the purpose of forecasting (*supervised - regression*) the daily sales of 1,115 Rossmann Drug Stores across Europe. The intention is to provide evidence that one particular algorithm is more effective than other methods for this purpose or for this task.

Data Link: https://www.kaggle.com/c/rossmann-store-sales/data - Script: https://goo.gl/iMmEoV

Dataset Description

During the initial phase of the project, a relevant portion of the time was allocated to the evaluation of the dataset and the dimensions of variables in the context we were operating. The preprocessing analysis has been conducted in R, an open source software, the purpose of the dataset is to predict the sales level and the various avenues of exploration are shown in the figures below.

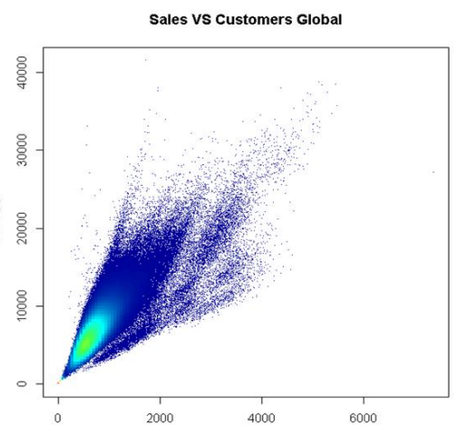


Chart 1

- Low correlation between the number of customers and volume of sales.

- We use this information to concur that **or** assess whether predicting the sales amount does not depend on the customers.

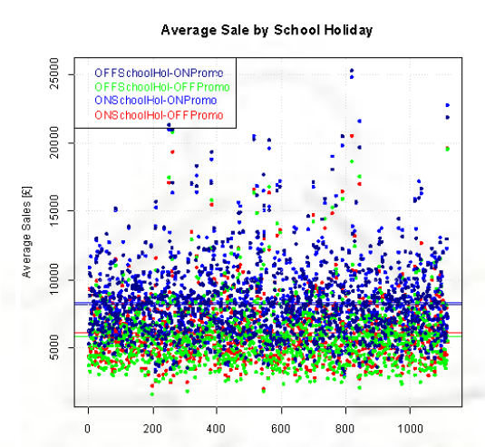


Chart 2

- School Holidays do not impact sales exclusively, when combined with Promotions there is a noticeable increases in sales.

- We considered this aspect while evaluating whether to consider school holidays as representative variable of the sales forecast we will perform.

- Substantial increase to sales when on promotion

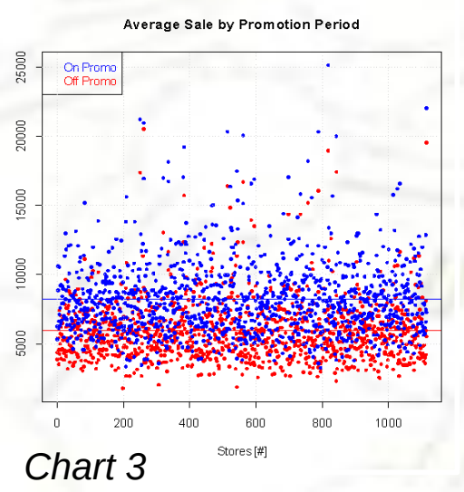


Chart 3

- Promotions only occur during weekdays - negligible impact on weekend sales.

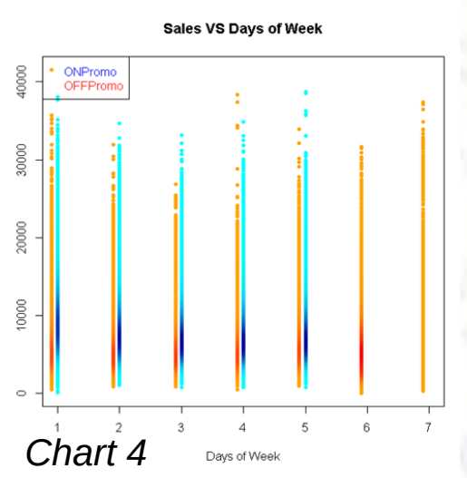


Chart 4

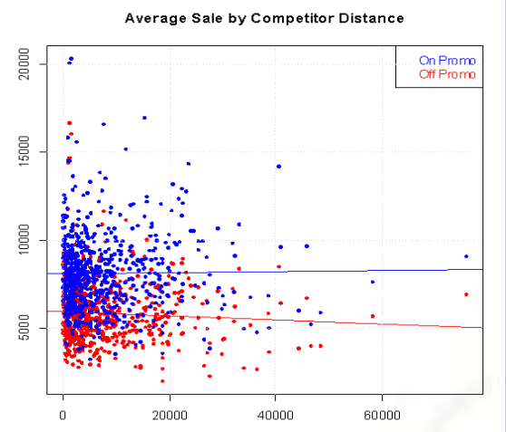


Chart 5

- Locality of nearest Competitor does not significantly impact sales.

- Data dictionary shown below for the first 8 features of the dataset

Variable	Mean	Std	Min	Median	Max
Store	558.43	321.91	1	558	1115
DayOfWeek	4	2	1	4	7
Sales	5773.82	3849.93	0	5744	41551
Customers	633.15	464.41	0	609	7388
Open	0.83	0.98	0	1	1
Promo	0.38	0.49	0	0	1
SchoolHoliday	0.18	0.38	0	0	1
DayInt	3	2	0	3	6
Weekend	0.28	0.45	0	0	1

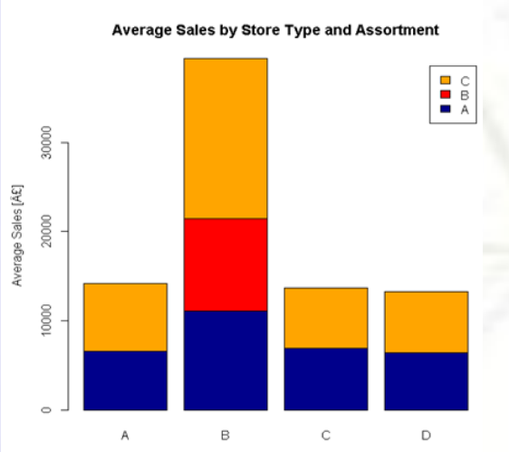


Chart 6

- Different store types are selling the same amount on average with the exception of stores with more assortments.

At this stage of the analysis and based on experimentation, it was decided during the selection of the bounds of the search space for Hyper-parameter optimization that the Feature Engineering would be beneficial to both models (lower MSE). The feature engineering that was created added an additional 24 features which were made up of Median Sales by Store for each Month by both Promotion and School holiday (*Median was chosen due to skewness*). Where the store had no respective value the mean (*Global*) value was taken (*this happened when there was no observation of either the store of event occurring in the training data). The Charts below (11-14) show the average number of Sales and Customers by Month, the most notable period is the lead up to Christmas.*

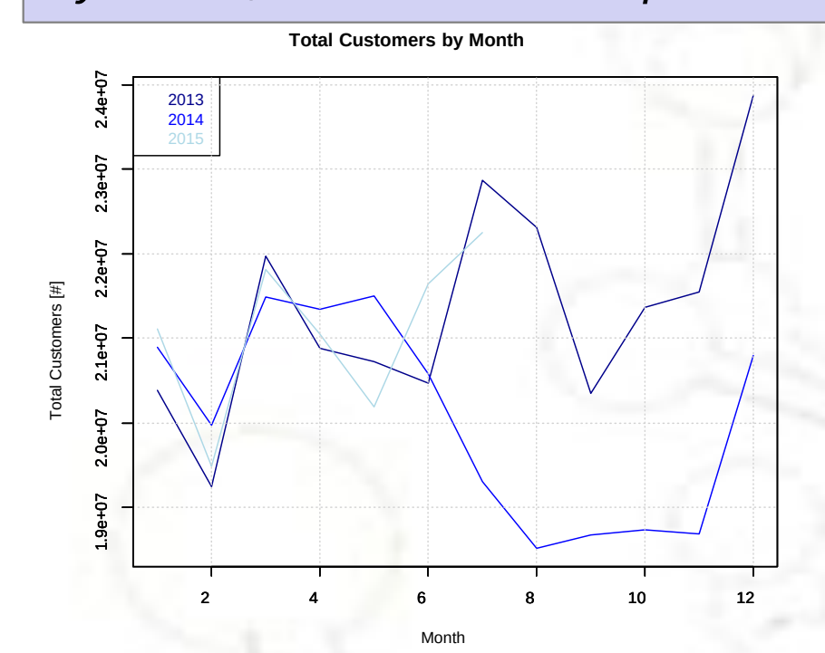


Chart 11

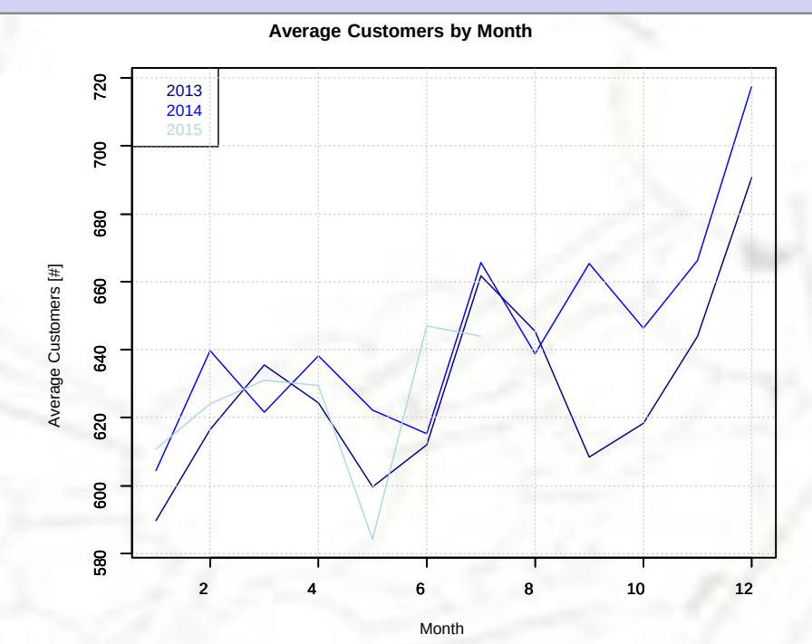


Chart 12

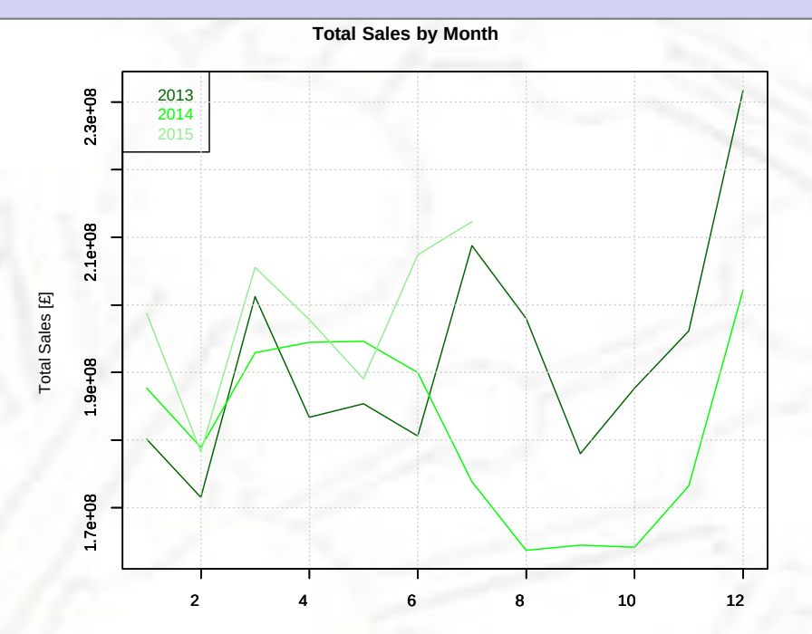


Chart 13

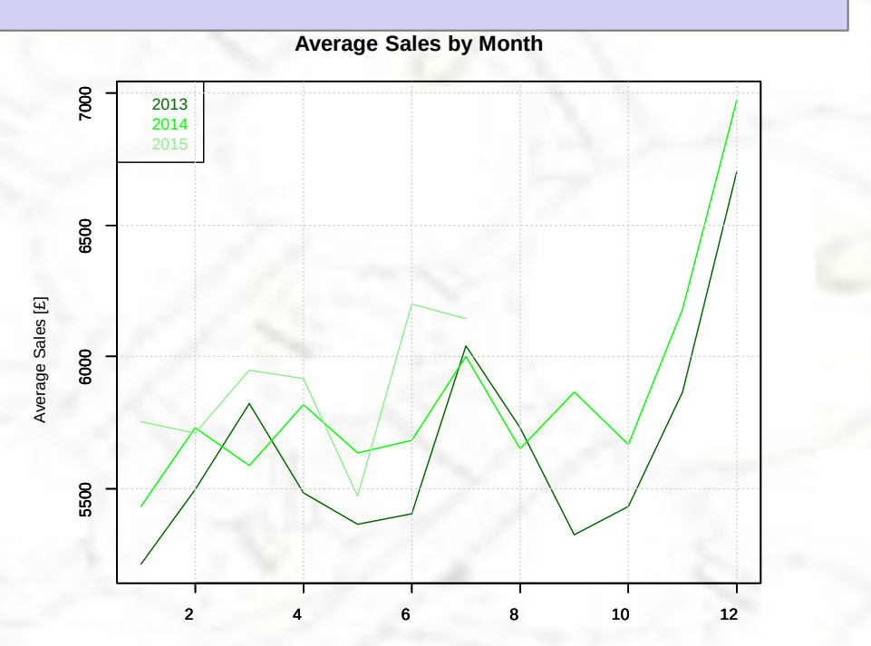


Chart 14

Hypothesis

Does a computationally more expensive, with regards to the number of models created, does Random Forest Machine Learning algorithm offer a better trade-off in terms of time to optimize and performance than K-Nearest Regression for the task of forecasting Sales for a entire Company's stores? Both we require extensive use of Optimisation in order to get optimal results.

Comparison of Two Algorithms

Algorithm	K-Nearest Neighbours	Random Forest
Summary	This algorithm works in a similar way to the clustering/ classification version, the variation of this algorithm means that it finds similar examples for each vector. Once the algorithm has summarised the data into *K* groups, predictions are made by taking average of each of the *K* groups as its prediction [**5**].	An ensemble method for both classification and regression type problems. The ensemble is built up by training many decision trees that when typically averaged can produce very good results.
Strengths	1. No assumptions about the data or characteristics are required. 2. Method is capable of dealing with non-parametric datasets. 3. Scalable, flexible and simple to implement.	1. The method for estimating missing data and maintaining accuracy when a large proportion of the data are missing is effective [**3**]. 2. It handles large set of input variables without variable deletion and thus without information loss. 3. Gives estimates of what variables are important in the regression.
Weaknesses	1. Finding k-nearest examples for examples can be expensive, complexity is of order *n* log(*k*). 2. Inference of the model is hard since there is no "model" description of each of the kernels. 3. Time consuming when K starts to get large.	1. The error rate depends on the correlation between any two trees in the forest, since increasing the correlation increases the forest error rate. 2. Overfitting of the data is a concern when the number of trees get large.

Description of Training and Evaluation Methodology

- Import the dataset
- Transformations, pre-processing and cleansing of the dataset
- Partitioning of the dataset; train and test
- Setup experimentation; Finding hyper-parameter bounds
- Setup of the training evaluation phase:
  - Hyper-parameter optimisation method; Bayesian
  - 75 Trials per model
  - 5 Fold Cross-Validation
  - Evaluation criteria; Mean-squared error [**2**]
- Interpret, analyse and evaluate training results
- Re-train models with optimal parameters
- Obtain predictions and evaluate

K-Neighbours Regression

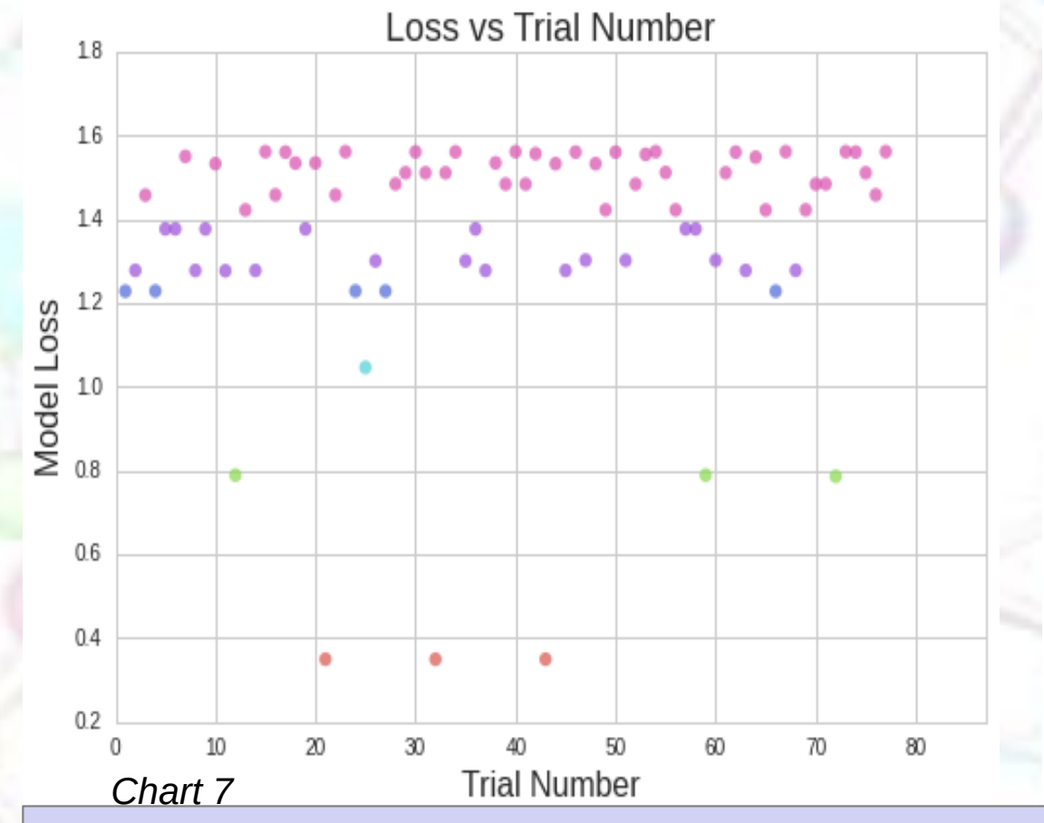


Chart 7

Random Forest

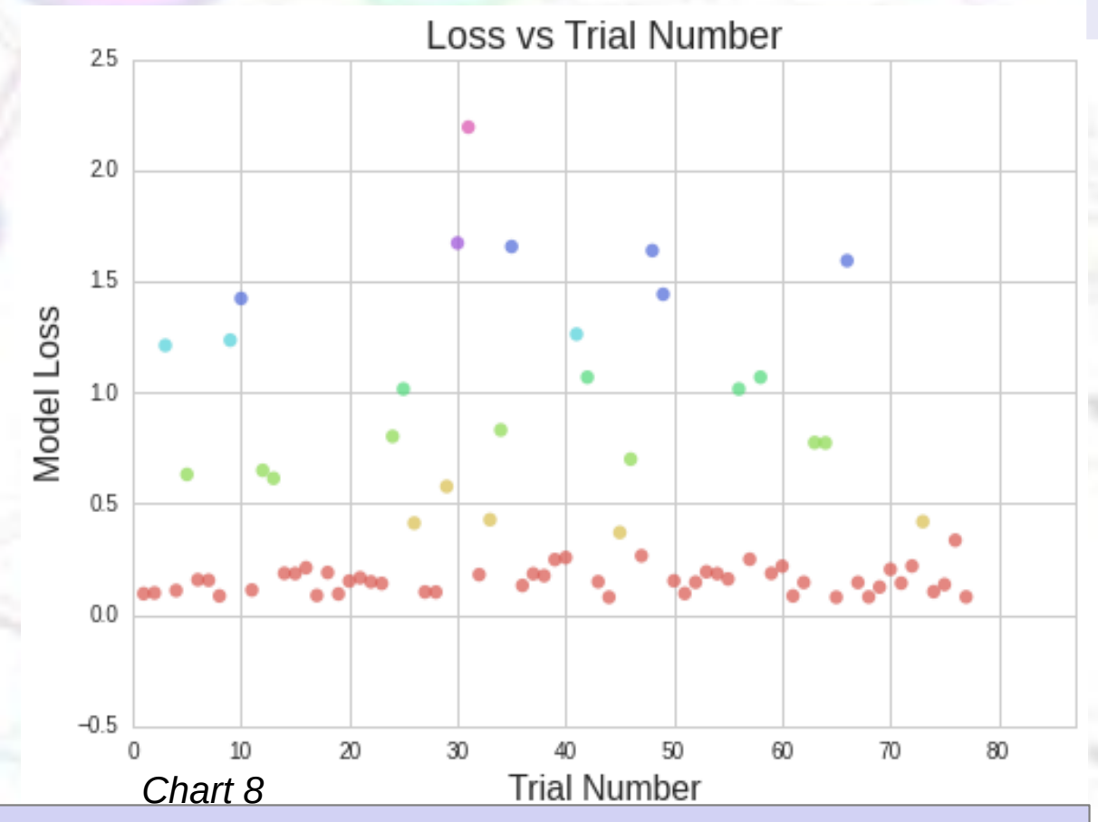


Chart 8

The MSE of the two models are compared respectively in the two charts above (*charts 7-8)*. It can seen that, in general, the Random Forest algorithm out-performs the K-Neighbours Regression with regards to their Mean Squared Error value. It must be noted that the errors in the K-Neighbours regression appear in horizontal bands in contrast to the Random Forest which is a lot more random.

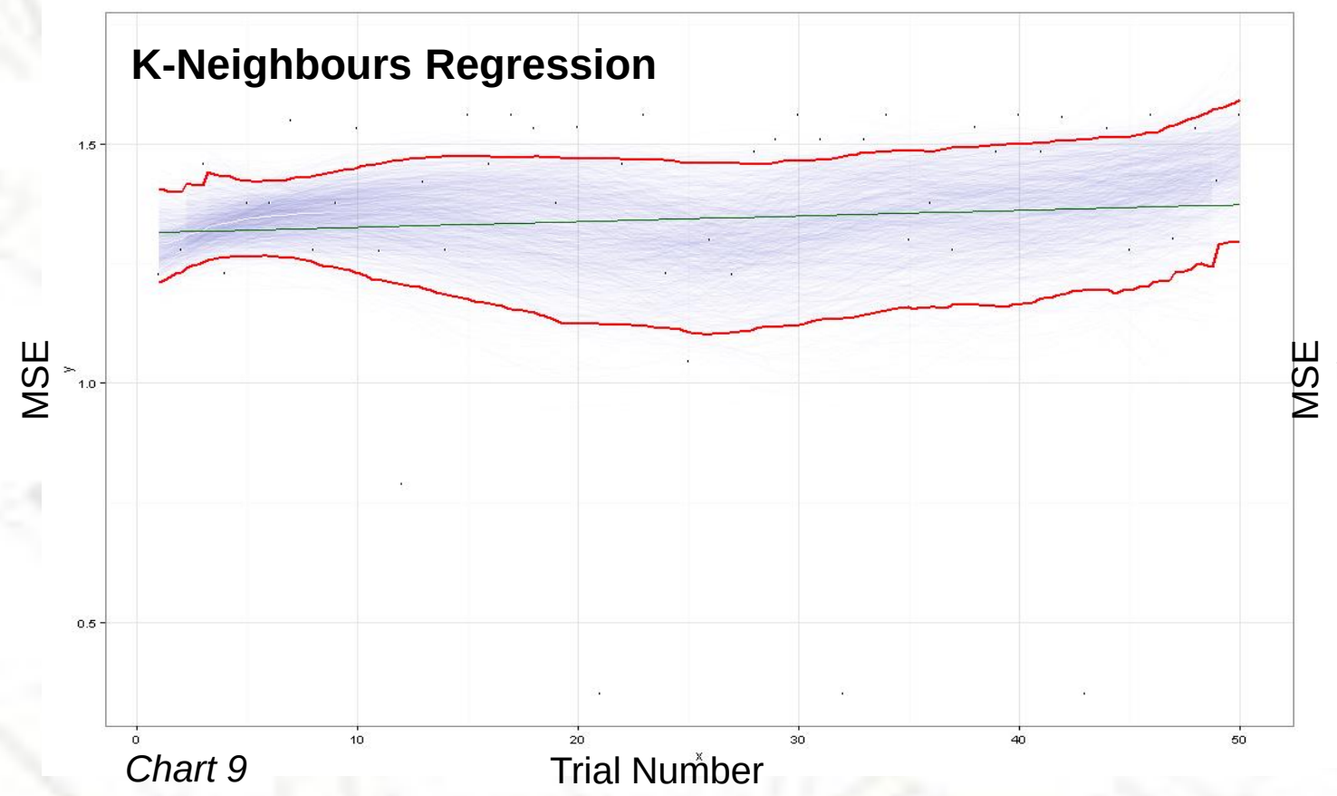


Chart 9

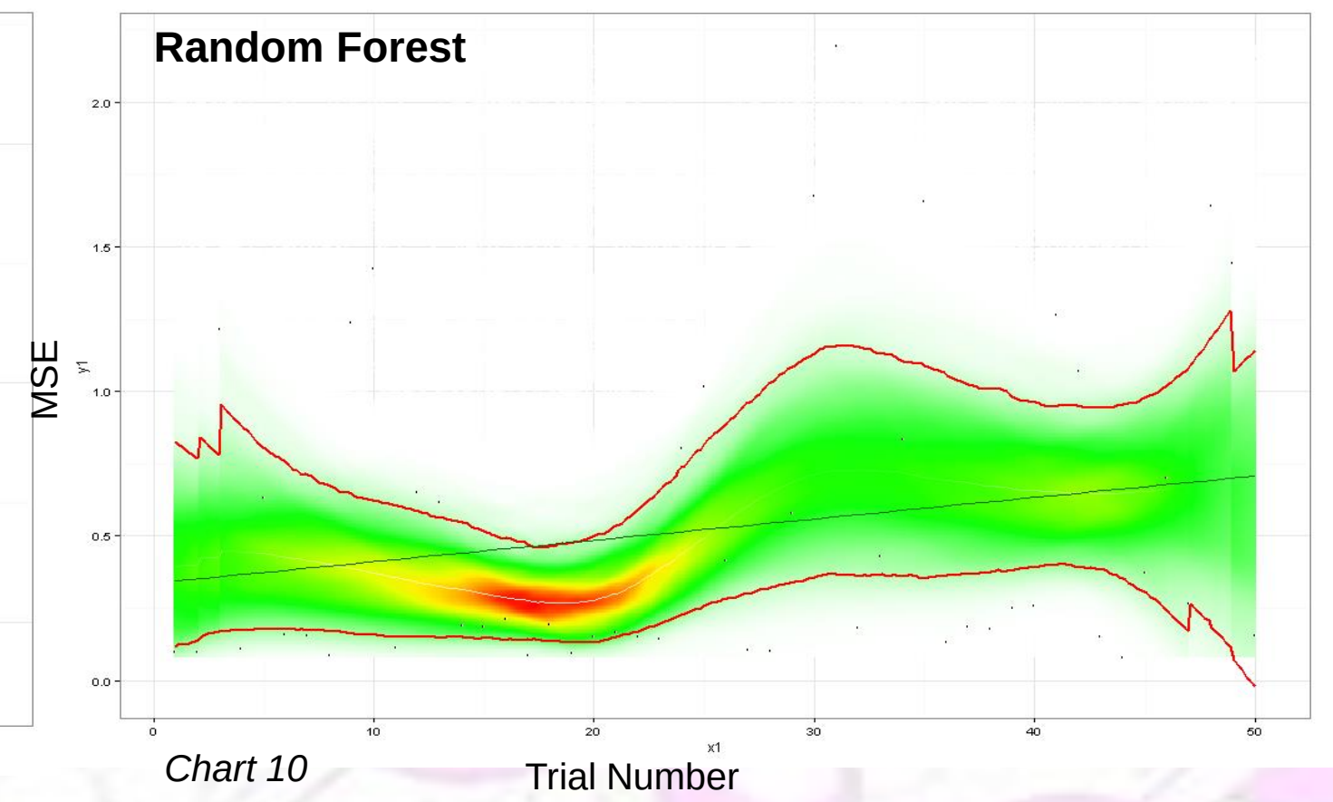


Chart 10

The two scatter plots provides a comparison of how the trials and the MSE are different. The KNN (Chart 9) provides less volatile MSE but starts at a higher average model loss. The Random Forest (Chart 10) instead produces lower MSE model losses on average but outliers are less predictable and are more volatile overall.

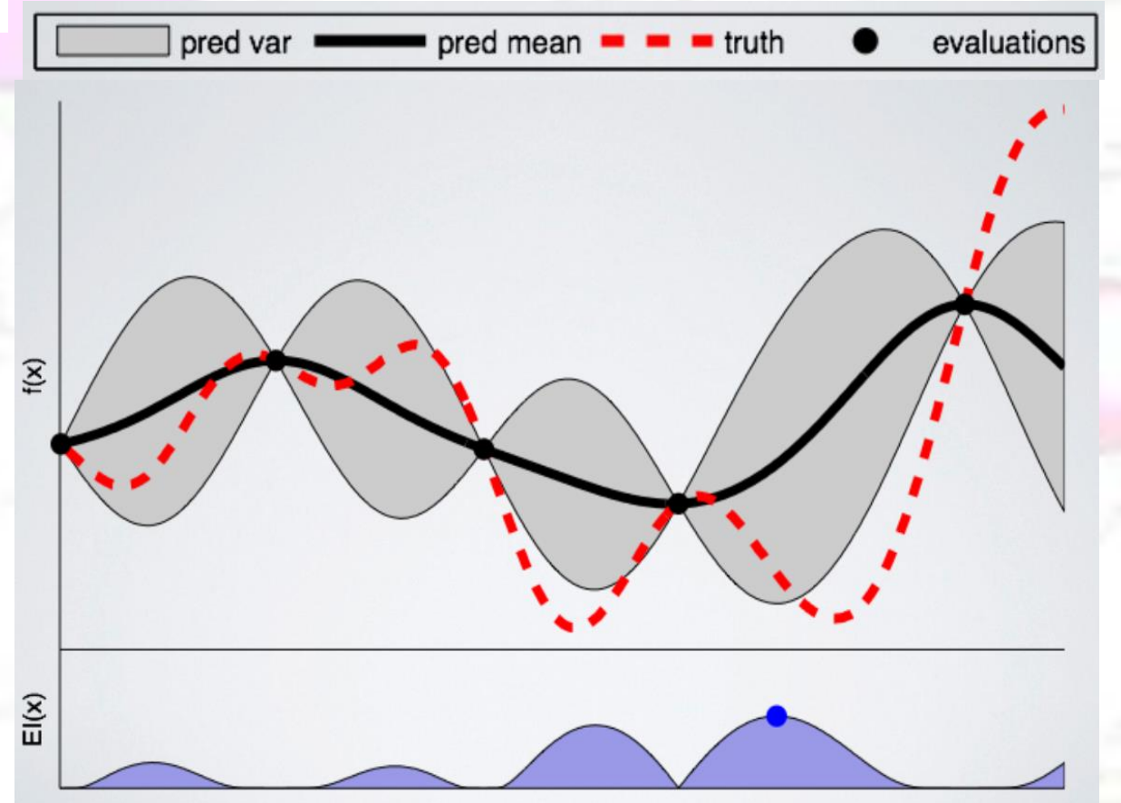
K-Fold Cross Validation

Is applied for two purposes, firstly to ensure that that the reported performance of an algorithm is as true as possible and secondly for the aid of model selection. The first point is crucial, a true representation of the average five samples of the training data is better than one at estimating the generalization error of a given model. The second point is that the output of each of the trials (*mean value of the cost function*) is used for determining the most appropriate model for the given task, forecasting the sales of the stores.

Hyper-parameter Optimization

Bayesian Optimization [*BO*] works in a probabilistic manner by exploring the search space of a set of hyper-parameters, the method strives to find and exploit areas with a low mean and high variance (*Blue Point on Figure 1*). The possible points that can be used are calculated based on their expected probability of improvement, this is shown in the figure to the right where evaluated points are denoted by the black points. Also in the figure, the blue point is where the maximum expected value of improvement is the greatest, this would therefore be the next evaluated point.

Figure 1 – Illustration of Bayesian Optimisation in Action



Bayesian Optimisation was originally developed in the 1970's by Jonas Mockus [**4**] as a technique for finding Global maxima of black-box type functions. This method offers an alternative to other well know methods of Parameter Optimisation, via the tuning of the models hyper-parameters, such as Grid Search or Random Search.

Choice of Parameters

Prior to running the full experiment, an exploratory analysis was undertaken to aid with determining the best hyper-parameters [*HP*] to work with. The output of this analysis are reflected in the Experiential Results section charts.

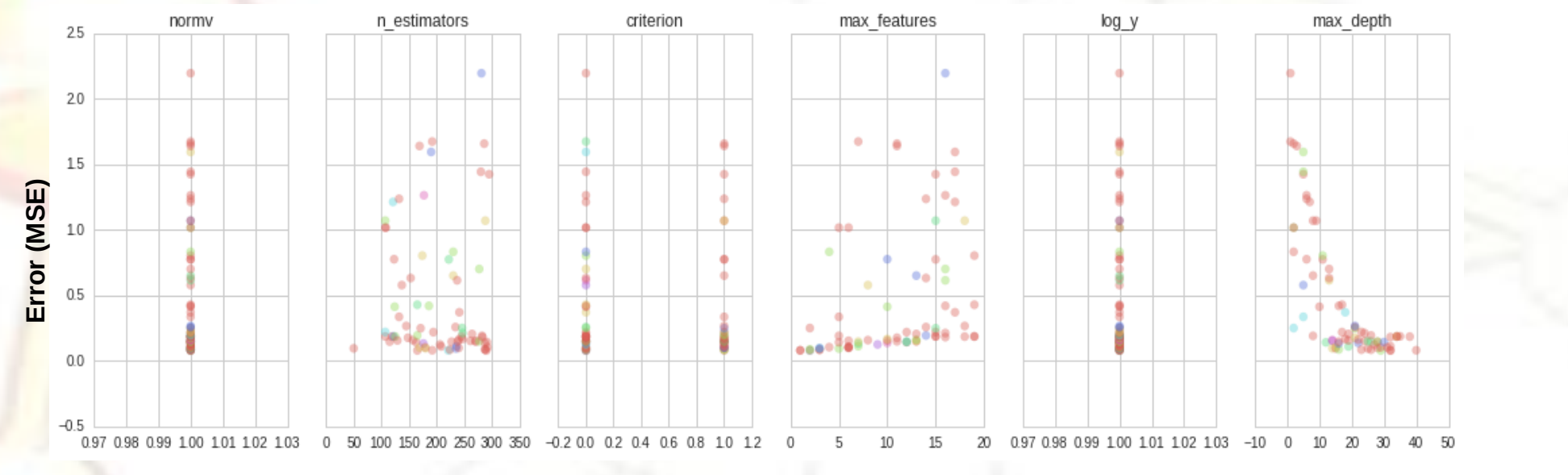
There were a number of HP that were experimented with; normalization and scaling of the datasets, logarithm of the target value and also the hyper parameters of each respective model.

HYPER PARAMETER	Search space for each Algorithm		Optimal Results for each Algorithm	
Algorithm	K-Neighbours	Random Forest	K-Neighbours	Random Forest
Max Features	-	(1-20)	-	2
Max Depth	-	(1-40)	-	29
Number of trees	-	(100,300)	-	290
Number of Neighbours	(2,20)	-	3	-
Leaf Size	(10,60)	-	56	-

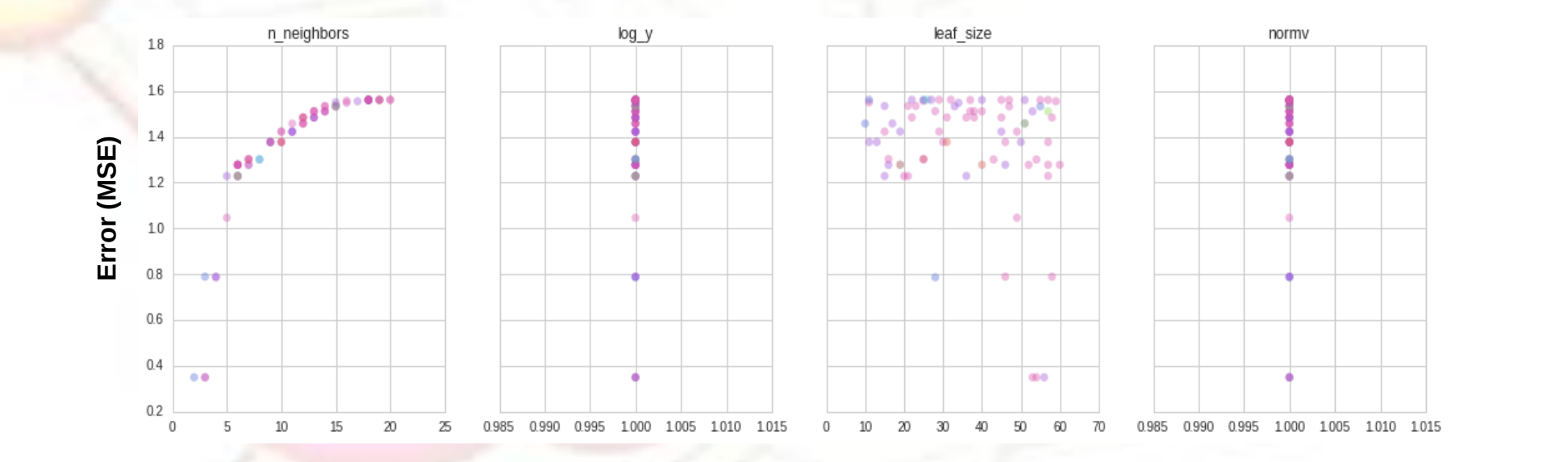
Experimental Results

Model loss is lower in the Ensemble model, in contrast to the KNN method where the models loss was consistently the same in respective of the HP configuration. There are varying HP combinations resulting in similar performance value in the Random Forest method and consistent in the KNN.

Random Forest Parameters by Error



K-Neighbours Regression Parameters by Error



Analysis and Evaluation

The model complexity of a ensemble method far exceeds the learning capability and performance of a single KNN regression model, this can be observed in the both the charts; Predicted Sales (*Chart 15*) and the plot of errors from each model (*Chart 16*)

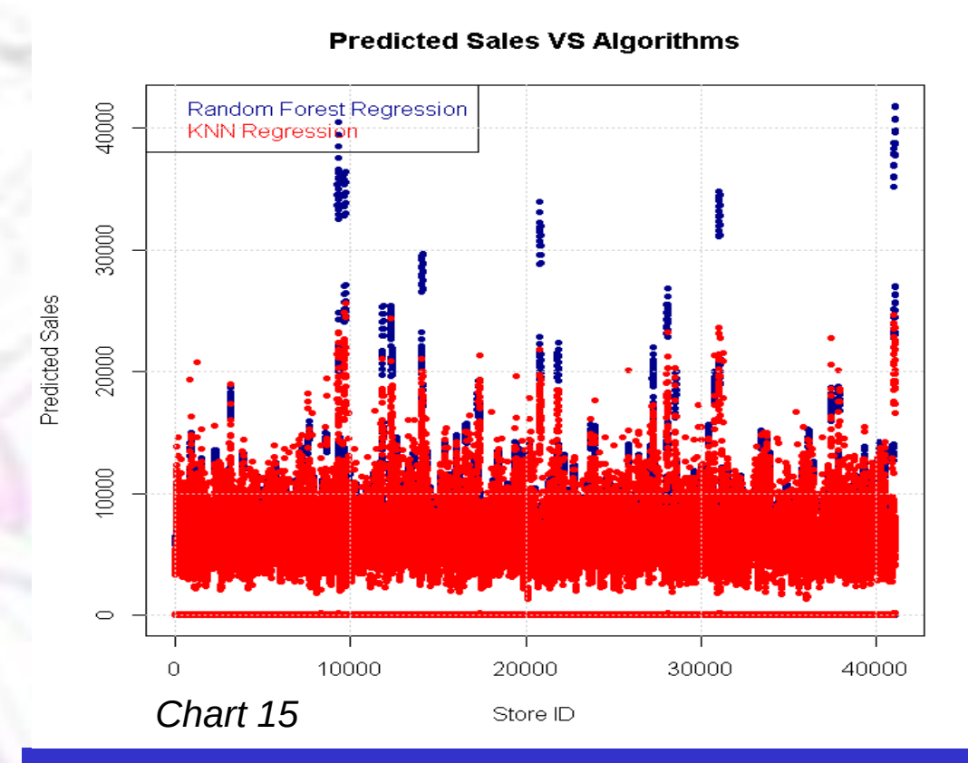


Chart 15

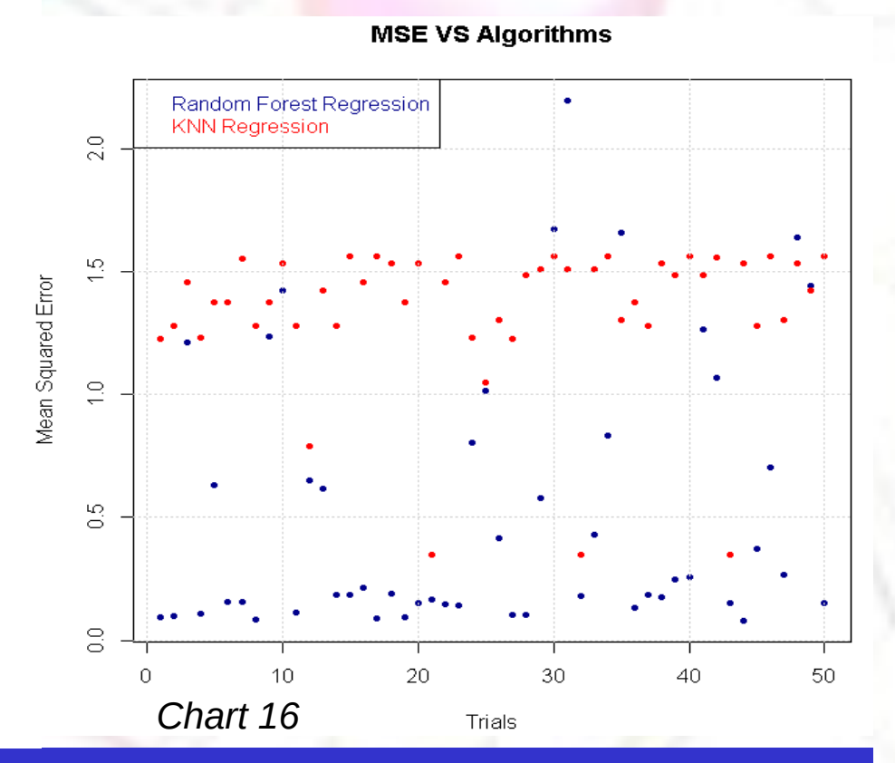


Chart 16

Algorithm	Test Set Results
K Nearest Regression	0.28703
Random Forest	0.11563

For this analysis, the test set was separated from the training set and had not been evaluated on until the optimal models had been found. The final results are shown above. The most notable evaluation is that the Random Forest method generalises substantially better than the K Nearest Regression algorithm. Likely reasons have been listed in the summary table, but in addition to those weaknesses the Kernels (*Neighbours*) are just too smooth for the given data and do not capture the variation of all the stores well enough.

Lessons Learned

1. Ensemble methods, specifically Random Forests, from the analysis deal well with a mixture of categorical and numerical features. When using hyper-parameter tuning in a multidimensional environment it is difficult to determine if the number of trials was sufficient and also if the optimal solution has be found.
2. Finding suitable Algorithms that can handle large datasets within the constraints of the system memory was a concern, this meant that the alternative algorithm to Random Forest had to be changed a couple of times to find a suitable one for the comparison.

Future Work

The underlying and pertinent property of the data is of a time series nature, therefore future work might include addressing time series analysis techniques; for instance trends, seasonality and also more advanced properties namely the impact and lag properties of sales. Additionally, exploring the use of Ensemble methods, the changing of combination rules (*voting based method used in this analysis*) and finally experimentation with building an ensemble by combining a variety of different Machine Learning models together in one ensemble.

References

[1] - Barton, D., and Court, D. 2012. “Making Advanced Analytics Work for You.” Harvard Business Review 90:79–83.
[2] - Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research 30, 79.
[3] - L. Breiman. Random Forests. Machine Learning, 45:5–32, 2001.
[4] - Jonas Mockus: On Bayesian Methods for Seeking the Extremum. Optimization Techniques 1974: 400-404
[5] - Yao, Z., Ruzzo, W.L., 2006. A Regression-based K nearest neighbour algorithm for gene function prediction from heterogeneous data. BMC Bioinformatics 7, S11. doi:10.1186/1471-2105-7-S1-S11